The manuscript presents an application of an ensemble-based physical data assimilation technique to a global biogeochemical ocean model, with a focus on the effect of physical data assimilation on climate-relevant carbon estimates. The manuscript is mostly well written and offers some valuable insights on the effects of physical DA, but the text could be improved in places and several aspects of the DA experiments should be examined further.

general comments

One aspect that is becoming more important in modeling studies but is seemingly ignored in the current version of the manuscript is the reporting of model uncertainty -- even though ensembles are used to generate the results. The authors mention ranges of estimates when reporting results from other studies. However, in their own analysis, the focus is solely on the ensemble mean, without examining the full model ensemble or reporting any uncertainty estimates. It would be beneficial to explore ensemble-based ranges of estimates and compare them to the improvements brought about by data assimilation. This could lead to interesting questions, such as the extent to which data assimilation constrains estimates and whether the estimates improve in areas where they are more constrained. Additionally, figures like Fig. 4 and the seasonal difference plots could be enhanced by including uncertainty estimates, such as the ensemble standard deviation or the interquartile range.

Thank you for the suggestion. Indeed, a reduction in the uncertainty of the CO_2 flux estimate would be a very relevant result in addition to an improved estimate of the mean CO_2 flux.

There is, however, one difficulty in the interpretation of the ensemble standard deviation in our method. Because we use a Kalman filter variant, the ensemble standard deviation (STD) of the DA-updated variables (T, S, SSH, u, v) is reduced in ASML. Most of the reduction in ensemble spread occurs over the course of the first year. After that, the STD remains stable, precisely because we tune our ensemble perturbation and ensemble inflation in such a way that the STD of temperature is maintained after the initial phase (Figure R1; yellow and green lines).

It is thus expected that the ensemble standard deviation of CO_2 flux decreases as well in ASML, but this is a result of the model and not part of the tuning. Indeed, we find that the STD for the local CO_2 fluxes in ASML is reduced to about 75-80% of the STD in FREE after the first year of assimilation (see example in Figure R2; however, this data is not area-weighted). We will add analysis and discussion of the uncertainty estimates in the revised manuscript in a computationally efficient way (rerunning for additional output is needed and may be done for one or more years).

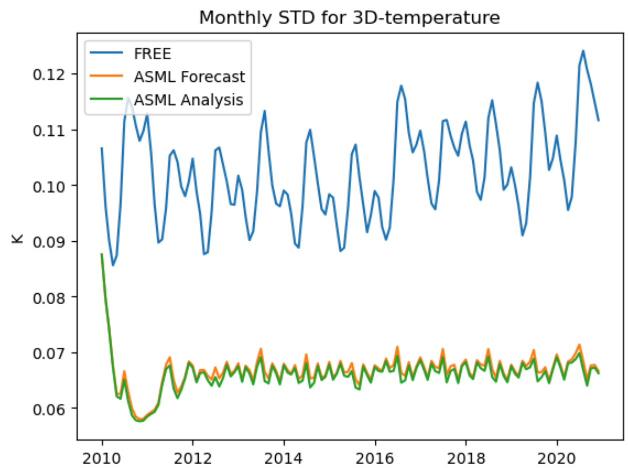

Figure R1: Ensemble standard deviation for 3D-temperature. Note: No volume-weighting applied for the global mean (includes empty cells).

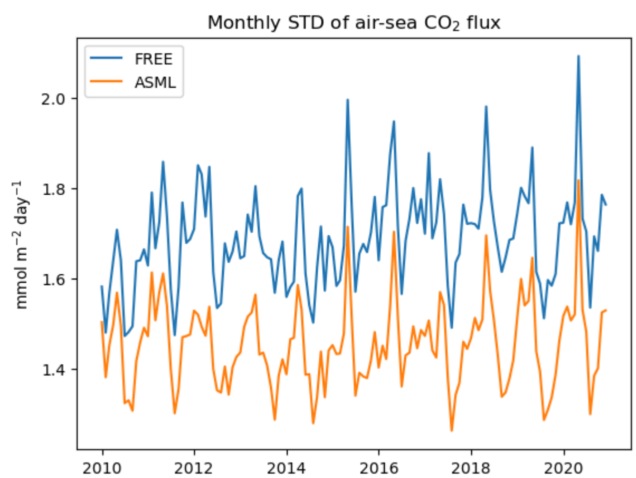

Figure R2: Ensemble standard deviation for the CO2 flux. Note: No area-weighting applied for the global mean.

Additional output with ensemble statistics for the year 2020 suggests that there are again regional differences. For example, in the Newfoundland Basin, which showed a strong effect of DA on CO_2 fluxes, the standard deviation was reduced strongly by assimilation, but less in the other regions in the North Atlantic (Figure R3). Discussion of these effects will be added to the revised manuscript.

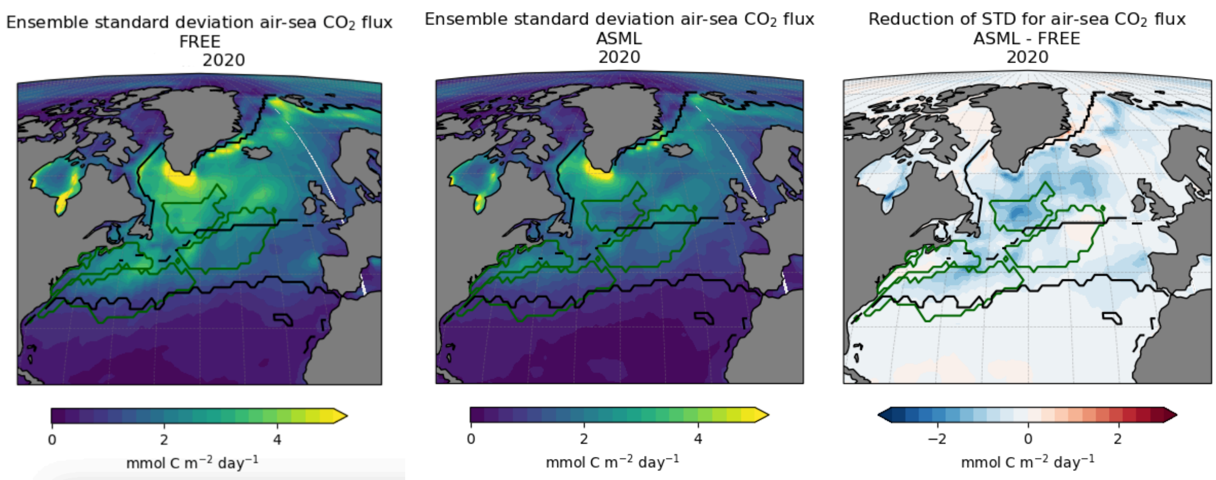

Figure R3: Ensemble standard deviation of CO_2 flux and its reduction for the year 2020 (last year of the simulation).

The manuscript emphasizes carbon storage through physical transport, i.e. "upwelling and subduction of DIC, as well as the physical transport of other biogeochemical tracers" (l 60). However, the role of biological carbon fixation and sinking of particulate organic matter seems underexplored. Given that the model includes both slow and fast sinking detritus variables, a more comprehensive examination of these processes would be valuable. Here, it would help to clarify whether the biological carbon export at 200m (l 379 and following) is primarily due to sinking or physical transport. A closer examination or clearer description of the effects of the DA on the biological drivers of carbon export would help to improve the manuscript.

We would like to note that we're most interested in anthropogenic CO_2 uptake, which is primarily physically driven (e.g., Gruber et al., 2023 https://doi.org/10.1038/s43017-022-00381-x). A much closer examination of the biological carbon pump would be interesting, but is beyond the scope of this paper. Yet, on a regional scale, changes in biological export production contribute to the overall carbon balance and thus may have noticeable effects on the regional net CO_2 fluxes.

In response to the reviewer's comment, in the revised manuscript, we will address explicitly where the assimilation produces a change in export production. We will provide supplementary maps of carbon export through sinking of detritus at 190m. For example, we find that in the North Atlantic Central STSS, the increase in export production (by up to 4 mmol C m^{-2} day^{-1}), presumably in response to mixed layer deepening and/or increase of SST, is essential to explain the overall effect in the direction of more CO_2 uptake (Figure R4).

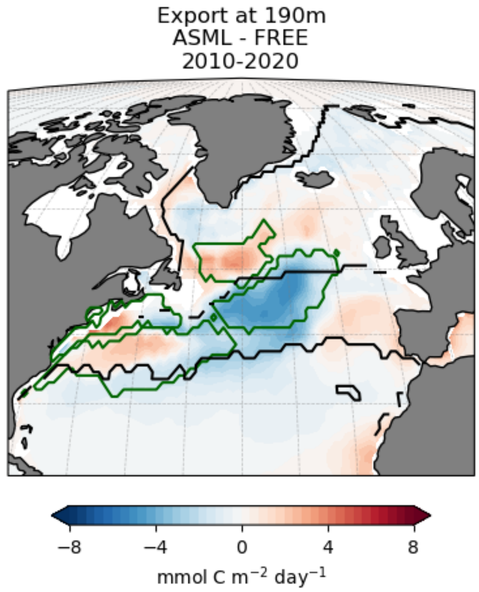

Figure R4: Effect of assimilation on carbon export through sinking of detritus at 190m.

In contrast, in the Southern Ocean, assimilation-induced changes in the air-sea CO_2 flux driven by physics (the effect of SST on pCO_2 and transport of DIC and alkalinity) are about twice as large as the response of the biological pump to the assimilation. In particular south of 50°S, the response of the biological pump is more than compensated for. The export of carbon through sinking of detritus decreases presumably because of shallower mixing. However, this is outweighed by a decrease of upward DIC transport and thus more ocean CO_2 uptake (Figure R5).

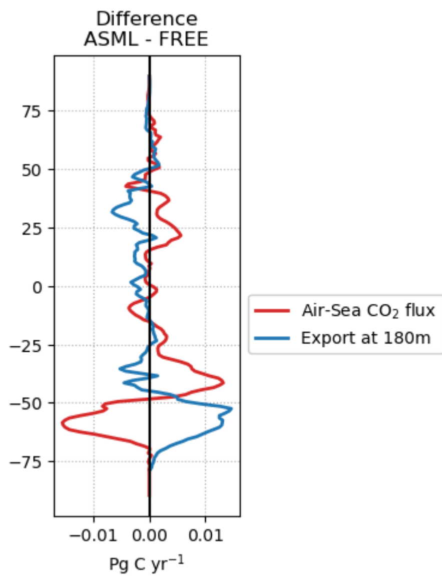

Figure R5: Assimilation-induced changes in the air-sea CO_2 flux and in carbon export through sinking of detritus by latitude. Negative denotes a more downward flux, i.e. CO_2 flux from air to sea and downward sinking of detritus.

We would also like to clarify that vertical export of organic carbon takes place almost entirely through sinking (the gravitational pump; Boyd et al., 2019: https://doi.org/10.1038/s41586-019-1098-2) rather than through physical transport. To clarify, we have changed these lines to:

(l. 379) “biological carbon export at 200~m through sinking of detritus”.
(l. 60) “physical transport of DIC ~~of other biogeochemical tracers~~” (deleted)

This is because the carbon biomass itself is very low: In REcoM, the integrated global ocean living carbon biomass is around 2 PgC and the mass of dead organic carbon is of similar magnitude. However, the biological fluxes are very efficient so that globally, sinking of detritus removes about 10 PgC (i.e. a multiple of its own mass) per year from the upper 200m of the ocean. Due to the low concentration of carbon biomass, transport by advection and mixing of organic mass is less important.

In contrast, the concentration of DIC in the ocean is much higher. In REcoM, the ocean globally holds around 38000 PgC in the form of DIC. Therefore, the transport of DIC by advection and mixing plays a decisive role.

The assimilation of physical observations that only directly updates the physical variables can lead to "shocks" in the biogeochemical variables. It would be valuable to know if the authors observed any negative effects of daily physical updates on the biogeochemical state, such as unexpected phytoplankton blooms (for example, caused by a deepening of the mixed layer transporting nutrients, formerly below the mixed layer, to the surface).

We are not aware of any such shocks. This might relate to our overall finding that the modeled carbon fluxes and other inspected variables such as chlorophyll-a, NPP and plankton biomass act almost surprisingly indifferent to substantial differences in the model physics. The most rapid assimilation-induced changes take place in the first few months after the start of the assimilation, yet there was no noticeable shock.

Several aspects of the model setup and data assimilation process could benefit from further explanation or discussion. For instance, the restoration of surface salinity towards climatology may interfere with the assimilation of salinity data. It would be informative to know if the authors have experimented with switching off the nudging when or where salinity data is being assimilated, and how well the salinity climatology aligns with the assimilated data.

The main effects of SSS assimilation and salinity restoring are to reduce SSS globally. In addition, there are certain regions of model bias, such as the Amazonas river inflow area and the North Atlantic Current, where both methods are consistent with each other. While there are gaps in the SSS-CCI data near the poles, the salinity restoring towards climatology is with global coverage. Experiments with and without salinity restoring show that without restoring, sea surface salinity in FREE drifts by approximately +0.05 psu during the first year after switching it off. In ASML, the difference between switching salinity restoring off or on is smaller (less than 0.01 psu globally), because the assimilation compensates for the lack of restoring. In ASML, global SSS is reduced by approximately 0.15 or 0.2 psu, respectively, after one year, which shows that the assimilation has a stronger effect than the restoring. The best agreement with SSS-CCI observations is achieved when assimilation and salinity restoring are used simultaneously.

In summary, we added to the manuscript: “Additional experiments with and without salinity restoring towards climatology show that the best agreement with the SSS-CCI observations is achieved by simultaneously using assimilation and restoring. Hereby, a benefit of additional restoring is the global coverage of the SSS climatology.”

Similarly, the exclusion of temperature observations from the DA when the model-observation difference exceeds 2.4°C could use a better explanation, as this seems to hinder assimilation where it might be most needed.

By excluding these observations, the aim is to prevent strong and sudden corrections from making the model unstable, especially in the initial phase. Instead, a ‘gentler’ correction is made by assimilating neighboring points. Because we use a gap-filled SST observational product, observations are continuously available in the neighboring domains. After the initial phase, about 7% of SST observations are excluded because of the temperature-threshold regularly. However, the data assimilation still has a strong effect in areas where these large model-observation discrepancies are found (North Atlantic Current, near Japan and in some places of the Southern Ocean).

We added this information to the manuscript.

To improve readability, particularly for readers less familiar with data assimilation techniques and carbon modeling, brief explanations of key concepts and modeling choices would be beneficial. These would include descriptions of the term used to perturb atmospheric forcing, the role of ensemble inflation, and the rationale behind the choice of γ_DIC and γ_Alk in Equations 4 and 5 (see also my specific comments below). Currently, the manuscript often uses references to other studies to motivate implementation details, and an additional sentence here and there could help the reader to better understand these details without having to go through other papers.

We agree and have made text additions in the places that you mentioned.

In places, the structure of the manuscript can be improved to enhance clarity and flow. Sections 4.2 and 4.3 are quite lengthy and could be subdivided based on location (Southern Ocean, Atlantic) and the different data products used in the comparisons. Section 3, which contains results from the two ensemble simulations, could be merged with Section 4 to create a more cohesive results section.

Thank you, we rearranged the sections and section titles accordingly.

Overall the figures look very good and are helpful, I only have a minor suggestion here: it might be more informative to report ASML-OBS instead of ASML-FREE in Figures 1-3. This would provide a clearer picture of the model error following data assimilation. Also, some of the figures, such as Figure 7, have lots of whitespace that could be reduced.

We have chosen ASML - FREE throughout the manuscript because it allows us to visualize comparatively small changes in some of the biogeochemical variables. On the one hand, for temperature and salinity, ASML-OBS provides a clear picture of the model error after data assimilation (see SST, Figure R6). On the other hand, for the biogeochemical variables, FREE-OBS and ASML-OBS are visually too similar to recognize the differences (see chlorophyll, Figure R7). Showing ASML-FREE for all variables allows one to recognize correlations between the effects of DA on different variables.

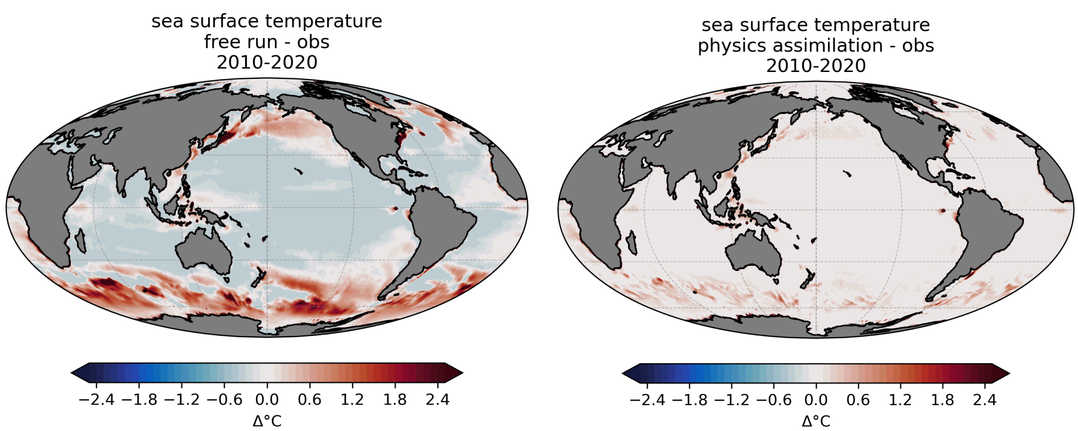

Figure R6: FREE-OBS and AMSL-OBS for SST, useful to illustrate the model error before and after assimilation.

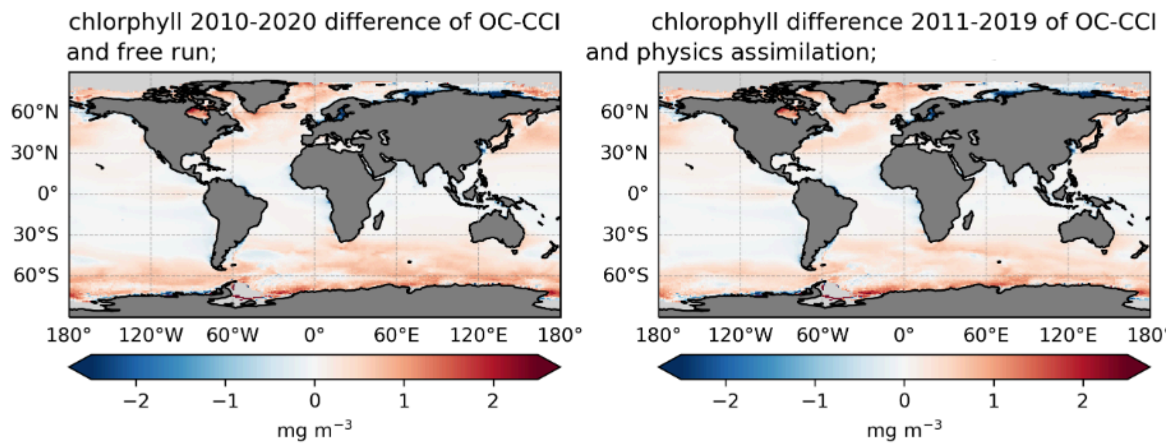

Figure R7: FREE-OBS and AMSL-OBS for chlorophyll, not useful because the effect of the assimilation is almost unrecognizable.

specific comments

L 8: "the mean CO2 uptake increases by 0.18 Pg C yr−1": Add "regionally" here to make it explicit that this increase is not a resulting global estimate.

Thank you. Done.

L 40: "the model mean": It would be helpful to the reader to add a few words about the kind of models that were considered here.

On this, added: “the mean of GOBMs included in the Global Carbon Project”

L 65: "DIC" was used before the abbreviation is introduced here (l 59). The earlier sentence actually makes a quite similar point about subduction of DIC and also mentions upwelling, perhaps this could be made more concise.

Thanks, we merged both sentences into one.

L 65: "It was shown that assimilating ocean physics at the initial state of a model simulation has a stronger and more positive impact on the modeled carbon cycle than assimilating the BGC initial state": Is this due to the lack of BGC observations mentioned earlier, the importance of physical processes for carbon export, or a large physical model error that cannot be decreased through BGC DA?

This study (Fransner et al., 2020) relates the strong and positive effect of assimilating ocean physics to the strong control ocean physics exerts on the biogeochemical variability on interannual to decadal time scales (rather than low availability of BGC observations or strong physical model errors).

The next sentence brings up the question of which processes are most important. Maybe a few candidates could be named and briefly discussed here before going into the details of the DA algorithm.

We name some candidates in the revised text:

“This raises the question which mechanisms produce the response of the \ce{CO2} flux in physics DA approaches. Is it the transport of DIC and alkalinity through physical advection, mixing and in particular upwelling of carbon-rich waters, as the model velocities and diffusities are changed by the assimilation? How much is \ce{pCO2} changed directly through its temperature-dependence? Does the biological pump respond to the assimilation of physics? How large are these effects, and when and where do they occur?”

L 70: "continuously assimilating ocean-physics for eleven years": A bit more detail could be useful here as well: What does assimilating ocean physics entail, what observations are being used for the DA here?

Thank you, added: “We continuously assimilate temperature and salinity observations for eleven years and update the modeled temperature, salinity, horizontal velocities and sea surface height.”

Detailed info can be found in the methods section.

L 89: "The model allows for a variable mesh resolution": What is a typical coarse and fine resolution used in the model grid?

Please see section “Simulation set-up”, which we have now moved here:
“The mesh resolution is nominally 1 degree, ranging between 120km and 20km with enhanced resolution in the equatorial belt and north of $\SI{50}{\degree N}$ (126858 surface nodes).”

L 93: A salinity flux of 0.1m/day? Please describe this better.

Thanks for asking. This number was a typo. We corrected the number and added description:

“The surface salinity is restored towards the World Ocean Atlas climatology through a fictional surface flux with a velocity of 50m /300 days according to the equation:

$(S_{\mathrm{clim}}-S_{\mathrm{model}}) / \mathrm{layer_width} * \mathrm{velocity}$”

For the example of a salinity bias of 0.5 psu and with the surface layer width being 5m, this would yield a correction of 0.016 psu per day.

L 96: "DIC" is introduced again, a quick search shows 7 introductions of "DIC", also counting captions.

Thanks, we only kept it in the Introduction and Conclusion.

L 117: "observations are weighted by distance": This is not a precise statement that could confuse some readers, express more clearly that the ensemble estimated correlation between a model grid point and an observation is down-weighted using a distance-based metric. Is vertical localization applied as well?

Thank you for the more clear wording suggestion, we used it. And we added that there is no vertical localization.

L 124: It would be useful to add equation numbers to all equations, even those that are not referenced in the text, so that they can be more easily referenced in other texts, such as this one.

Done.

Eq L 124: Why does a larger ensemble amplify rand? It does not seem that intuitive to have larger perturbations in a larger ensemble.

The incomplete definition of ‘rand’ in the manuscript has led to an obvious misunderstanding: In fact, there are no larger perturbations in a larger ensemble. The factor (N_ens-1) compensates that the values of ‘rand’ become smaller with increasing ensemble size. The values for rand are generated by Second-Order Exact Sampling from a trajectory of atmospheric forcing fields, a method introduced by Pham et al., see e.g.:
https://doi.org/10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2
and briefly explained here:
https://pdaf.awi.de/trac/wiki/EnsembleGeneration

To clarify, in the updated manuscript we add a few sentences on the generation of the initial perturbation and the stochastic element ‘rand’.

L 153: "model values are computed as the average of the grid points of the triangle enclosing the observation because the number of observations is fewer than model grid points": Averaging is required to interpolate the model solution at the observation locations, why is this dependent on the number of observations?

Thanks for pointing out how this can lead to confusion. In fact, we simply meant: If observations of a variable are spatially highly resolved, they are interpolated to the model grid (as for SST and SSS). If observations of a variable are sparse, it is the other way round and the model solution is interpolated to the observation locations (as for the profile data).

As this was unnecessarily confusing, we have now removed the last part of the sentence.

L 157: This information about the model grid is missing from Section 2.1 where the model grid is described for the first time. It would also be useful to describe the atmospheric forcing before describing the perturbation to it (Section 2.2.1).

Thank you, we have moved the section.

L 171 "the river flux adjustment (...) is applied to the pCO2 products. ...": It is not entirely clear what this means, the focus here is just the CO2 flux associated with the oceans, I presume? The next sentence provides some more information but it seems to imply that the RECCAP2 CO2 flux is not being used for comparison, when previous sentences stated that it was. Some clearer language would be useful here.

In response to both reviewers pointing this out, we have rephrased:

“We present \ce{CO2} flux estimates for the period 2010-2020, that are compared to the 'Regional Carbon Cycle Assessment and Processes 2' (RECCAP2) global air-sea \ce{CO2} flux estimates \citep{devries2023}. To make the RECCAP2 estimates comparable with our estimate stemming from a model without river carbon input, we apply a river flux adjustment \citep{friedlingstein2023,regnier2022} to the RECCAP \ce{pCO2} products. Thus, we quantify the anthropogenic perturbation of the ocean carbon sink \citep[as $\mathrm{S_{OCEAN}}$ in the Global Carbon Budget][]{friedlingstein2023,hauck2020}, and not the contemporary net air-sea \ce{CO2} flux with outgassing of river carbon into the atmosphere (as in RECCAP2).”

L 183: Should the US East Coast be considered subpolar, are all regions characterized by seasonal stratification, or does SPSS stand for something different here? A alternative choice of region names may be suitable and would avoid confusion with the region names in the Southern Ocean.

In the revised text we point out that, according to the definition of Fay and McKinley, the STSS, SPSS and ICE biomes exist analogously in both hemispheres (https://doi.org/10.5194/essd-6-273-2014). Therefore, there is an SPSS biome in the North Atlantic, of which we discuss only specific parts (e.g. the East Coast SPSS region).

The Fay and McKinley biomes are used widely in the ocean carbon cycle community (see e.g. RECCAP papers, https://reccap2-ocean.github.io/publications/).

To avoid confusion with the regions names in the North Atlantic (NA) and Southern Ocean (SO), we will add subscripts the biomes, e.g. $\mathrm{STSS_{SO}}$+ and East Coast $\mathrm{SPSS_{NA}}$

L 185: Please explain "NAC".

North Atlantic Current, done.

Eq 1 and 2: Is there an easy to communicate motivation for the choice of γ_DIC and γ_Alk ?

In order to assess the dynamic DA effects on surface \ce{pCO2}, it is useful to distinguish between different variables that constitute the change in \ce{pCO2}. Oceanic \ce{pCO2} varies mainly with temperature, DIC and alkalinity. Thus, we decompose changes in \ce{pCO2} into their contributions from changes in SST (SST), surface DIC and alkalinity (Alk). For that, we apply the following approximations of \citet{sarmiento2006} and \citet{takahashi1993}:

[equations]

Here, differences between ASML and FREE are denoted by Δ; else, the average of ASML and FREE is used for the computation. The sensitivities γ_{DIC} and γ_{Alk} describe how \ce{pCO2} varies with changes in one variable while keeping the other variables constant. For the sensitivities, we use an approximation derived from the solution chemistry of carbon dioxide in seawater following \citet{sarmiento2006}:

[equations]

Eq 1, 2 and 3: Previously Delta denoted the difference between ASML and FREE, is this still the case here? If so, are the regular terms (e.g. DIC in Eq 1 or the terms in γ_{DIC}) from the FREE experiment? This should be mentioned in the description.

Yes, delta is the difference between ASML and FREE and the regular terms are calculated from the average of the two - this has been added above.

L 220: Why not mention EN4-OA earlier when the other data products are introduced?

Okay, done.

L 250: "at greater depth than 500 m, where the model's subsurface temperature": The "subsurface" can be deleted here.

Okay, done.

L 266: Please explain what a 15%-line is.

"The maximum extent of sea-ice in September, here defined as the area where the sea-ice concentration is more than 15\%, is smaller in FREE than OSI-SAF, which is demonstrated by the 15\%-line surrounding that area for FREE and OSI-SAF (\cref{fig:sic}a);"

L 301: "In the more northern part of the STSS, which we call the STSS+, the CO2 uptake is reduced ...": The text here could be considered misleading because STSS+ is not defined as the northern part of the STSS, but as the part of the STSS with a positive CO2 flux difference. I would prefer a change in formulation that avoids this ambiguity, for example: "The part of the STSS characterized by a positive CO2 flux difference between ASML and FREE, which we call the STSS+ and in which the CO2 uptake is reduced, forms an outer (northern) ring around the STSS region." The same comment applies to STSS+ a few lines below.

Thank you for the suggested wording! We made use of it.

L 373: "the effect of the DA is towards increased uptake of CO2 during boreal summer and autumn in ASML (Fig. 6g). This prevents summer outgassing": The increased summer uptake prevents summer outgassing, isn't this just describing the same effect? I would suggest rewording this sentence.

“In the Central STSS, the effect of the DA is overall towards an increased uptake of \ce{CO2} from May to November (\cref{fig:CO2_NA}g). In particular, this prevents outgassing in high summer and even reverses the flux direction for some months, so that there is uptake in ASML almost all year round (\cref{fig:CO2_NA}c).”

L 411: "(difference of FREE and SOCAT in (Fig. 9a); difference of ASML and SOCAT not shown)": The figure label claims that ASML - SOCAT is shown.

Thank you for noting this. Figure data and labels have been updated to show FREE - SOCAT, as indicated in the text.